Incorporating Single Arm Evidence into a Network Meta Analysis Using Aggregate Level Matching: Assessing the Impact

Joy Leahy[1,2], Howard Thom[3], Jeroen Jansen[4], Emma Gray[5], Aisling O'Leary[2] , Arthur White[1,2], and Cathal Walsh[2,6]

[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2]National Centre of Pharmacoeconomics, St. James Hospital, Dublin 8, Ireland
[3]Bristol Medical School: Population Health Sciences, University of Bristol, UK
[4]Stanford University School of Medicine, Department of Health Research and Policy Epidemiology, Stanford CA, USA
[5]School of Medicine, Trinity College Dublin, Ireland
[6]Health Research Institute and MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland

Abstract

Increasingly, single armed evidence is included in Health Technology Assessment submissions when companies are seeking reimbursement for new drugs. While it is recognised that Randomised Controlled Trials provide a higher standard of evidence, these are not available for many new agents which have been granted licences in recent years. Therefore, it is important to examine whether alternative strategies for assessing this evidence may be used.

In this work, we examine approaches to incorporating single armed evidence formally in the evaluation process. We consider matching aggregate level covariates to comparator arms or trials, and including this evidence in a Network Meta Analysis. We consider two methods of matching; 1. we include the chosen matched arm in the dataset itself as a comparator for the single arm trial, 2. we use the baseline odds of an event in a chosen matched trial, to use as a plug-in estimator for the single arm trial.

We illustrate that the syntheses of evidence resulting from such a setup is sensitive to the between study variability, formulation of the prior for the between design effect, weight given to the single arm evidence, and the extent of the bias in single armed evidence. We provide a flow chart for the process involved in such a synthesis, and highlight additional sensitivity analyses that should be carried out. This work was motivated by a Hepatitis C dataset where many agents have only been examined in single arm studies. We present the results of our methods applied to this dataset.

1. Introduction

A Network Meta Analysis (NMA)[1–8] allows researchers to formally combine all relevant direct and indirect evidence in order to give the best possible evidence to decision-makers. Randomised Controlled Trails (RCTs) are considered the gold standard of evidence, as their controlled approach minimises potential bias. However, sometimes not all treatments have been evaluated in an RCT, and the only available evidence on a treatment may be from sources such as observational studies or single-arm studies. These types of evidence may contain valuable information, especially when no other evidence is available, but they may potentially be biased ([9–11]). In a well conducted RCT we can be confident that the patients are exchangeable across treatment arms, as they have been randomly assigned. However, the same cannot be said for non-randomised evidence.

In one of the most recent reviews of its kind, Griffiths and Vadlamudi[12] examined submissions to three Health Technology Assessment (HTA) agencies between 2010 and 2015; NICE (UK), CADTH (Canada), and IQWiG (Germany). The proportion of HTA submissions that considered

non comparative evidence was 38%, 13%, and 12% for each agency respectively, although this may be a broader sample than simply single arm trials. Submissions based exclusively on non comparative evidence was 4%, 6%, and 4% respectively, making a total of 27 submissions, although some of these submissions may have included Individual Patient Data (IPD). Positive outcome rates for non comparative evidence alone versus overall submissions were 60% versus 84% for NICE, 69% versus 68% for CADTH, and 17% versus 38% for IQWiG. From this analysis it is clear that when RCTs are unavailable, HTA decision-makers are willing to consider non-comparative evidence, despite its limitations. Given that single arm evidence is being accepted by HTA agencies, it is essential to minimise potential bias by providing clear guidelines for incorporating this evidence into an NMA. In 2016 Bell et al[13] identified "priority research requirements" such as explorating the consequences of data being drawn from different settings for invention and its compator, and "the extent to which observational designs can complement or replace those of RCTs".

Matching-Adjusted Indirect Comparisons (MAIC)[14,15] or Simulated Treatment Comparisons (STC)[16,17] are emerging methods for reducing bias when incorporating single arm trials into an NMA. However these require IPD for at least some of the studies, and it is frequently the case that only aggregate data is available.

Much research has been undertaken on the inclusion of non-randomised evidence into an NMA to date. Sutton et al[18] investigate incorporating observational evidence in a pairwise meta analysis, while Schmitz et al[19] and Efthimiou et al[20] propose methods for including observational studies in an NMA in a manner that treats both evidence types separately. Thom et al[21] and Goring et al[22] investigate methods of including single arm evidence. Thom et al propose using a random effects model for the expected placebo effect in order to incorporate single arm evidence into an NMA. However, this interferes with the randomisation of the RCT evidence. Goring et al estimate absolute treatment effects and compare the results of their models to the recommended relative effects models. Other work on absolute versus relative effects includes Hong et al[23,24]. However, it has been argued by Dias et al[25] that absolute effects "effectively breaks randomisation, and in fact runs against the entire way in which randomised controlled trials are designed, analysed, and used".

Here we aim to address the research requirement proposed by Bell et al by assessing the appropriateness of including single arm evidence in an NMA through matching to other arms or trials with similar patient covariates. As we are dealing with single arm studies we match on both effect modifiers and prognostic variables, as recommended in Phillippo et al[26,27]. Firstly, we consider a method of choosing another arm from the network. This means that we treat the single arm and the chosen matched arm as if they are arms from the same study. We consider a naive pooled method and a more formal hierarchical model. The first approach was recently adopted by Jaff et al[28] in an application to assess the efficacy of endovascular interventions. They compare the NMAs which do and do not include single arm trials using matching. Schmitz et al[29] also use a pooled model to include single arm trials, and investigate the appropriateness of these models depending on how close the matches are to the single arm trial. Secondly,we match to a chosen trial by plugging in the baseline odds of an event in a chosen matched trial for the reference treatment. We compare these different models and evaluate the performance in a simulation study. We also recommend a number of sensitivity analyses which can be used to detect bias. This is illustrated in Section 4 by using an example from the treatment of Hepatitis C Virus (HCV) infection.

There are a number of advantages to our methods over the methods described above:

1. The method only requires aggregate data.
2. Although including matched evidence adds non randomised evidence into the network, the randomisation of the available RCTs is kept intact.
3. The methods fit within the relative effects framework.

4. We can apply this method to a network of any size.

The objectives of this paper are to:

1. Assess which parameters influence the accuracy of the model's estimate in an NMA;
2. Assess under what circumstances it is appropriate to include single arm evidence.

The remainder of this paper is organised as follows: A detailed description of the matched arm methodology is provided in Section 2. Choices of hyperparameters are also comprehensively discussed. Section 3 describes the construction and results of a comprehensive simulation study. Section 4 presents our methods applied to a HCV infection network. A general discussion and some recommendations are provided in Section 5.

2. Methods

2.1 *Model Development*

We construct a model where we observe r_{ij}, the number of events in the j^{th} arm of the i^{th} trial. Treatment 1 is considered the overall reference treatment, with all other treatments being compared to it. $r_{ij} \sim \text{Bin}(p_{ij}, n_{ij})$, where p_{ij} is the probability of an event in the j^{th} arm of the i^{th} trial and n_{ij} is the number of patients in the j^{th} arm of the i^{th} trial. n_{ij} is a fixed, observed quantity and p_{ij} is made up of δ_{ij}, the treatment effect in the j^{th} arm of the i^{th} trial and μ_i, the log odds of having an event in the baseline treatment (i.e. in arm one) for study i. We assume a random effects model, $\delta_{ij} \sim N(d_{t_{ij}} - d_{t_{i1}}, \sigma_\delta^2)$, where $d_{t_{ij}}$ denotes the effect of the treatment in the j^{th} arm of the i^{th} trial relative to the reference treatment and σ_δ represents the between trial standard deviation (SD) of the treatment effect. t_{ij} is the treatment in the j^{th} arm of the i^{th} trial, which we can refer to as treatment k. Let d_k denote the effect of treatment k relative to the reference treatment for the NMA as a whole. Note that $d_{t_{i1}}$ will only equal zero if the first treatment in a given study is the reference treatment for the NMA as a whole.

The model is written as:

$$\text{logit}(p_{ij}) = \begin{cases} \mu_i & \text{if j=1} \\ \mu_i + \delta_{ij} & \text{if j>1} \end{cases} \tag{1}$$

with priors $\mu_i \sim N(0, 1.83^2)$, $d_k \sim N(0, 1.83^2)$ and $\sigma_\delta \sim \text{Unif}(0, 2)$ in the case of two arms. We can adjust for trials with more than two arms by following Dias et al[30]. A table of the glossary of notation used throughout this paper is included in the supplementary material.

The hyper-parameters for $\boldsymbol{\mu}$ and $\mathbf{d}$ are chosen in order to have an approximate uniform distribution on the log odds ratio. Kass and Wasserman[31] point out that the properties of a prior on one scale can differ when transformed to another scale. A seemingly vague prior such as $N(0, 100^2)$ is not vague on the inverse logit scale, as most of the distribution is close to either 0 or 1. However, the choice of $\sigma = 1.83$ means that two standard deviations on each side of the mean covers 95% of the transformed (approximately uniform) distribution. This is illustrated in Leahy et al[32]. For the prior for σ_δ, when transformed to the probability scale, two standard deviations cover the range (0.02, 0.98), which we deemed to be sufficiently vague. This prior has also been examined by Lambert et al[33]. We compare the log odds ratios (LORs) for the RCT only dataset using a standard WinBugs prior versus the prior proposed in this paper. A table of the results is shown in the supplementary material.

2.2 *Including Single Arms*

We look at a number of ways for incorporating single arm trials into an NMA.

1. (a) Including the matched arm in data - pooled model: We treat the single arm trial and its chosen match as if they come from the same trial. In this case we group all evidence types together in such a way that different forms of evidence are not distinguished by the model. This is the most straight-forward model to implement and is set up as in Equation 5. In this case the number of studies is the total number of RCT and matched studies.
   (b) Including matched arm in data - hierarchical model: Again, we treat the single arm trial and its chosen match as if they come from the same trial. In this model we estimate the treatment effect, d, at each level of the study design, and then combine to give the overall treatment effect. This model provides more flexibility, at the cost of requiring a more stable network structure. The schematic of this model can be found in the supplementary material.
2. Plug-in estimator model: We assume the log odds of having an event on treatment 1 (reference treatment) is the same for the single arm trial and the chosen matched trial. This model has the advantage of not using any data more than once. Both the RCTs and the single arm trials are pooled as in model 1a above.

When using matching to incorporate single arm trials into an NMA, the main concern is that the chosen comparator will include patients from a very different population to the single arm trial. This could add bias to the model as one treatment could end up looking superior when it was simply allocated to a particularly healthy patient population. In order to minimise this potential bias we propose to choose matched comparators with the closest patient characteristics. When including a matched arm in the data (Model 1a or 1b), this match can be any other arm in the network, from either an RCT or another single arm trial, provided that the treatment is not the same as the treatment in the single arm to which we are matching. Let M be the number of covariates considered, let x_{m_k} be the proportion of patients possessing the characteristic associated with the covariate in the single arm trial with treatment k, and let $x_{m_{ij}}$ be the proportion of patients possessing the characteristic associated with the covariate in arm j of study i. The difference is: $\Delta_{ij,k} = \sum_{m=1}^{M} |x_{m_{ij}} - x_{m_k}|$. When matching to a trial using a plug-in estimator (Model 2) we can choose to match to any RCT, regardless of which treatment it contains. In this case the difference simplifies to: $\Delta_{i,k} = \sum_{m=1}^{M} |x_{m_i} - x_{m_k}|$. In our example, since we look at binary covariates, x is a proportion. However, this can easily extend to continuous covariates where x is the mean value of the covariate. The full steps and sensitivity analyses that can be carried out when matching single arms is presented in a flow diagram in Figure 1.

For the hierarchical model the extra level on the treatment effect is modeled as follows:

$$\begin{aligned} d_{\text{RCT}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}) \\ d_{\text{MATCHED}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}), \end{aligned} \tag{2}$$

where σ_{des} is the between study design SD which represents the variability between the RCT and matched studies. This is essentially a random effects model for the study design level. The prior distributions for d are the same as for the standard and pooled NMA models, and $\sigma_{\text{des}} \sim \text{unif}(0, 2)$. A benefit of the hierarchical model is that we can adjust for overprecision in the matched arms by applying a multiplicative factor ω to the matched precision, thus inflating the variance. This represents our increased uncertainty in the evidence from the matched arms, and can be thought of as the weight given to the matched evidence. For example, if ω is small this indicates that we believe that the matched evidence is a poor estimate of the mean effect.

$$\begin{aligned} d_{\text{RCT}[k]} &\sim N(d_{[k]}, \sigma^2_{\text{des}}) \\ d_{\text{MATCHED}[k]} &\sim N(d_{[k]}, \tfrac{\sigma^2_{\text{des}}}{\omega}) \end{aligned} \tag{3}$$

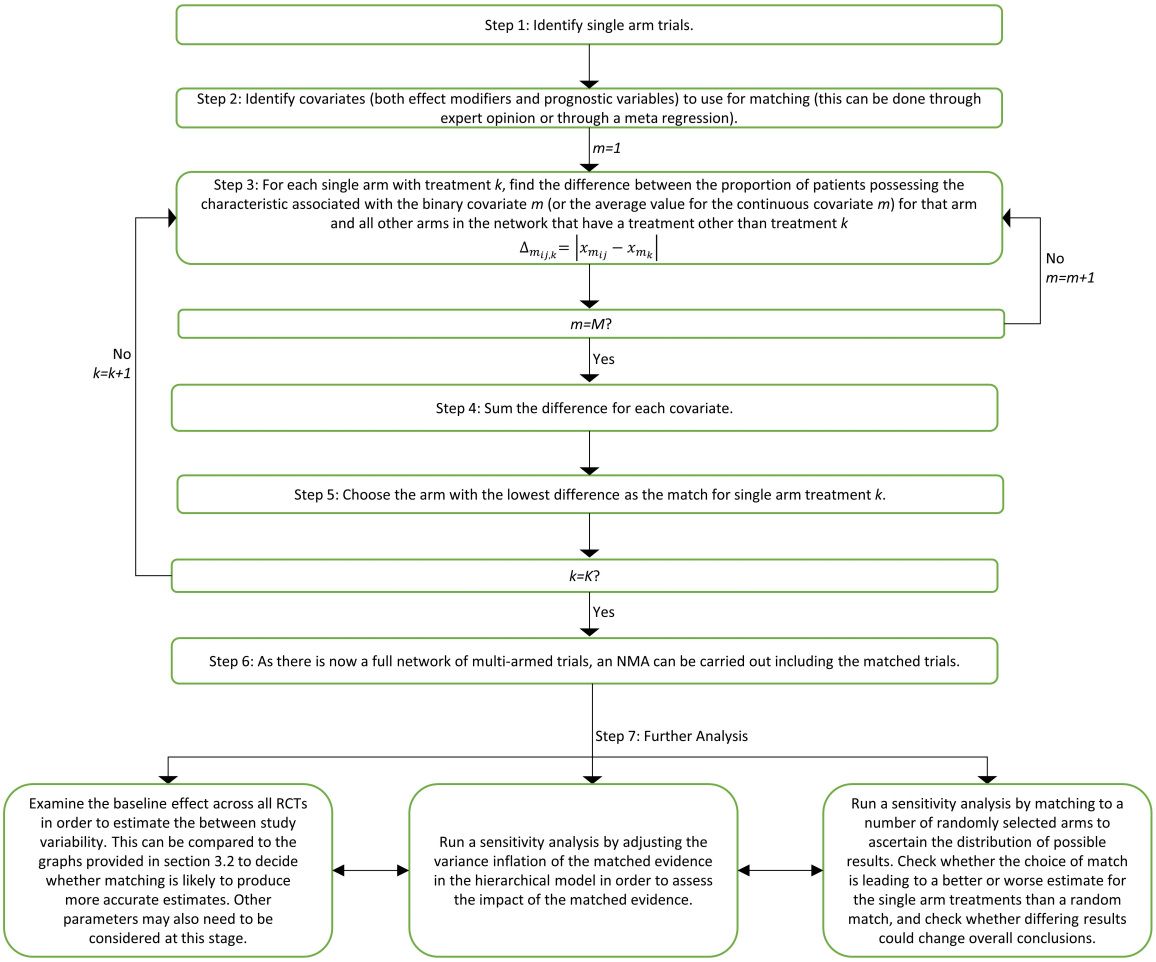


Figure 1: Steps to be taken when carrying out the matching by including an extra arm in the data, and recommended sensitivity analyses. $x_{m_{ij}}$ is the proportion possessing the characteristic associated with the binary covariate m (or mean of in the case of a continuous covariate m) in the j^{th} arm of the i^{th} trial, x_{m_k} is the proportion possessing the characteristic associated with the binary covariate m (or mean of the covariate in the case of a continuous covariate m) in the comparison arm, K is the total number of treatments examined in single-arm trials, and M is the total number of identified covariates.

Schmitz et al[19] investigate a number of different models for incorporating different types of evidence into an NMA and propose the hierarchical model as the best option, as we can obtain estimates for each study design levels, and down-weight certain types of evidence. However, due to the extra level in a hierarchical model, estimates may often be less certain and can be drawn closer to zero.

The plug-in estimator model is written for the RCT part as:

$$\text{logit}(p_{ij}) = \nu_i + \delta_{ij} \tag{4}$$

and the matched part is written as:

$$\text{logit}(p_l) = \nu_{\text{ChosenRCT[i]}} + \delta_l \tag{5}$$

where ν_i is the log odds of having an event for treatment 1 in trial i, and ChosenRCT[i] represents the chosen trial to match to single arm trial l. For the matched part of the model the j subscript is unnecessary on δ as this part includes single arm evidence only.

3. Simulation Study

3.1 *Methods*

3.1.1 *Set Up*

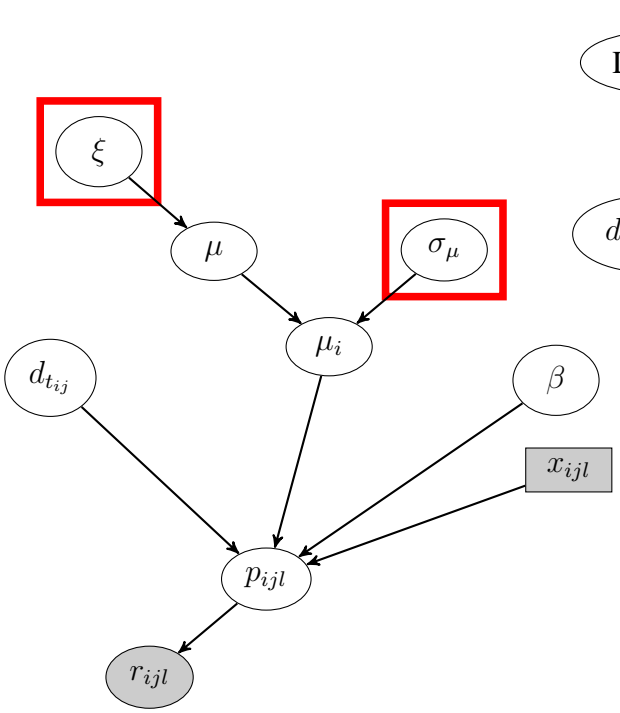
We compare the results of the matching methods (i.e. using the eight RCTs and the 4 matched RCTs) to only using the evidence from the eight RCTs. We also include a scenario in which we randomly choose a matched arm, in order to assess how much value there is in finding the best match. The parameters that are varied in the simulation study are described in Table 1 and are highlighted on the Directed Acyclic Graphs (DAGs) in Figure 2.

Table 1: Details of parameters varied in each scenario in the simulation study

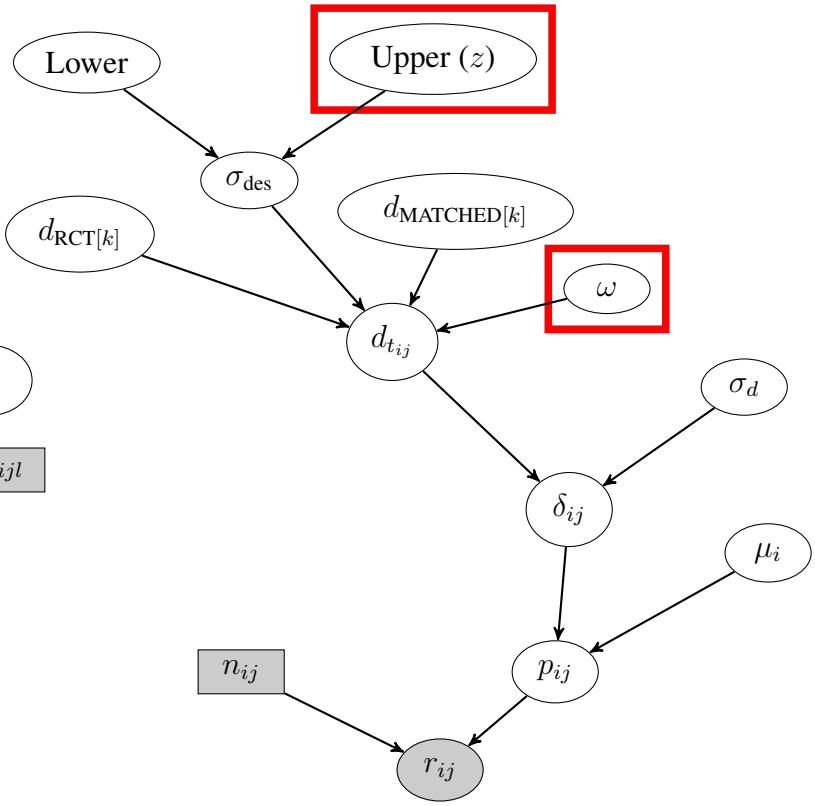
Parameter	**Description**	**Range**	**Notes**
σ_μ.	The SD of the baseline study effect. This is the measure of variability that will be used throughout this paper. It is the standard deviation of the logit of the baseline risk of having an event in each study in the network.	0-1.18	The highest value of 1.18 is approximately 3 times as large as the estimate of σ_μ in the HCV infection network. It should be noted that we chose to investigate this measure, instead of the more commonly investigated between-study heterogeneity, which quantifies how relative treatment effects varying between trials. When matching single arm trials we are interested in matching on prognostic variables and effect modifiers, whereas the between-study heterogeneity in RCTs is only affected by effect modifiers. Since σ_μ is driven by both unidentified prognostic variables and effect modifiers, we consequently use this as our measure of between study variability.
ξ	The bias in the single arm trials. The mean of the logit of the baseline risk of having an event in the RCTs comes from a fixed distribution $N(0, 0.59^2)$, while the mean of the logit of the baseline risk of having an event of the single arm studies varies between 0 and 1. In the most extreme case the single arm trials are simulated from $N(1, 0.59^2)$.	0-1	This corresponds to a baseline rate of 75% on the probability scale, so we believe that this is a sufficiently high value to use as the mean to cover plausible scenarios.
z	The upper bound of the uniform prior, $(0, z)$ on the between study design effect in the hierarchical model ,σ_{des}.	0.25-2	We believe 2 is a sufficiently vague upper bound as two standard deviations cover the range (0.02, 0.98) on the probability scale.
ω	The variance inflation on the matched evidence in the hierarchical model	0.1-1	This has the effect of downweighting the matched evidence.

When examining the hierarchical model, in order to examine the impact of model misspecification, two further scenarios were considered:

- Extra RCTs: When examining the prior on the between study design effect, σ_{des}, we also included a scenario with 12 RCT studies, but we artificially group eight on one side of the



(a) DAG for the simulation study. The parameters that are varied are dependent on the data themselves, and the investigator has no input. Note: ξ is 0 for RCTs.



(b) DAG for Hierarchical Model. The parameters varied are modeling choices which are chosen by the investigator. Note: This DAG simplifies to the RCT or Pooled Model by excluding all hyper-parameters on $d_{t_{ij}}$.

Figure 2: Directed Acyclic Graphs for simulations and models. The red box around certain nodes indicates that these are varied in a simulation study. Square nodes represent fixed quantities while the ellipses are stochastic nodes. The shaded nodes represent observed quantities.

hierarchical model, and four on the other side. This is to quantify the effect of the prior on the hierarchical model versus the effects of genuine difference in study types.

- RCT By Omega: When examining the variance inflation of the matched evidence in the hierarchical model, ω, we also include ω acting on the RCTs instead of the randomly matched trials, $d_{\text{RCT}[k]} \sim N(d_{[k]}, \frac{\sigma^2_{\text{des}}}{\omega})$.

A network of 12 studies with seven treatments was simulated. The studies consisted of eight two-armed RCTs and four single arm studies. The RCT studies consisted of treatments numbered 1-5 and the single arm studies consisted of treatments 1, 2, 6 and 7. To ensure our results were applicable to a wide range of real world networks we kept our network as generic as possible, and randomly assigned treatments to the RCTs at each iteration. Additionally, the treatment effects, study effects and covariate effects were simulated at each iteration. Further details on the network structure can be found in the supplementary material.

Data was simulated as follows:

1. We simulated:
    - the study effect μ_i;
    - the treatment effect $d_{t_{ij}}$;

- the effect of one baseline binary covariate β;
- x_{ijl}, an indicator variable representing whether or not each patient has the covariate.

2. We computed the logit probability of an event for each individual patient l, in arm j of study i: $\text{logit}(p_{ijl}) = \mu_i + d_{t_{ij}} + \beta x_{ij}$.
3. r_{ijl} was simulated from Bernoulli(p_{ijl}) for each patient l.
4. As we examining aggreagate data in this paper, we aggregated r_{ijl} to give an event rate for each arm, r_{ij}. In the case of continuous covariates our summary would be the mean, which takes the place of the proportion in the model. Thus our approach can be extended in a natural fashion to continuous covariates.

3.1.2 *Default Values*

The default values were chosen based on the HCV infection network in Section 4 or based on vague prior distributions. Each parameter was simulated independently. The values were set as follows:

- The between study variability, $\mu_i \sim N(0, 0.59^2)$. The default of 0.59 is half way through the range on which we are varying the between study variability.
- The relative difference of treatments to baseline, $d_k \sim N(0, 1.83^2)$, which is a broad range for this parameter.
- The number of patients for single arm trials was simulated from $n_i \sim \text{Unif}(75, 134)$, which reflected the inter-quartile range of the size of trials in the HCV infection network. The number of patients in the RCTs was twice this value.
- The probability of patients possessing each covariate was sampled from Unif$(0, 1)$ for each trial in order to cover the full range of possibilities. From this, each individual patient having the covariate was sampled from a Bernoulli distribution with the trial probability of having the covariate. This ensured that the RCT trials reflected the real world situation where treatment arms are exchangeable.
- The covariate effect size, $\beta = -1.04$, was set to be twice as large as the largest estimated covariate effect we found when analysising the RCT studies through a meta regression[34] in our HCV infection example, and was therefore thought to be adequately large to reflect possible real world covariates.

3.1.3 *Implementation*

We ran the models as described in Section 2 to assess how well they predicted the true relative treatment effects, d. Models were run using Markov chain Monte Carlo (MCMC) simulation in OpenBugs[35]. A burn in of 20 000 iterations was tested for convergence using the Gelman-Rubin statistic[36]. Following this another 10 000 iterations were sampled for our estimates. If the convergence condition was not met the number of iterations was doubled (both for the burn in and for the samples for estimation) and then tested again until the Gelman-Rubin statistic was less than 1.1. If the chains had not converged after a burn in of 320 000 this iteration was excluded from the analysis. Given the amount of simulations required we decided it was not feasible to include chains that had not converged by this point. If the chains did not converge for one of the models in a particular simulation the results for all other methods in the simulation were excluded from the analysis in order to eliminate any potential bias due to differing simulations. We varied each parameter in turn and sampled 8-12 data points for each at least 200 times.

In order to assess whether including single arm trials produces more accurate estimates than using RCT evidence alone, we look at the mean absolute error (MAE) for treatments 2-5: MAE $=$

$\sum_{s=1}^{T}\left\{\sum_{k=2}^{5}|d_{k_s}-\hat{d}_{k_s}|/4\right\}/T$, where $\hat{d}_k$ is our model's estimate of the relative effect of treatment k, compared to treatment 1, and T is the number of simulations. A lower MAE means that the model produces a more accurate estimate of the treatment effect.

We also compare the posterior SD as reported in the BUGS output to quantify the uncertainty in the resulting estimates. While we might expect a lower posterior SD when more evidence is added into the model, this might not necessarily hold due to the potential bias from the single arm studies. Graphs showing the MAE and BUGS posterior SD are included in the results section.

3.2 *Results*

The graphs in this section show the lines of best fit for the simulated data points, obtained by regression using both a linear and a quadratic term. Graphs showing the original data and the Monte Carlo Error (MC Error) of the simulations are included in the supplementary material. For simplicity, we use a pooled model for all analyses that do not directly investigate the nature of the hierarchical model.

3.2.1 *Parameters considered in exploratory analysis*

In an exploratory analysis a number of parameters were examined through the simulation study. We analysed the magnitude of the covariate and found that for larger covariate effects there was an increased advantage of choosing a matched arm based on covariates over a randomly chosen match. However, the magnitude of the covariate effect chosen for the simulation study was based on how large we would reasonably expect a single standardised covariate to possibly be. Heterogeneity between treatment effects was also examined. When there was large heterogeneity all methods were less accurate and less precise at estimating the treatment effect. However the loss of accuracy and precision was more pronounced for the RCT only evidence than when matched evidence was included. Finally, we examined how trial size affected the accuracy of treatment effect estimation. We found that including matched evidence was most beneficial when trials were small. For the remainder of this Section we will focus on the parameters highlighed in Section 3.1.1.

3.2.2 *Between Study Variability*

We examine the effect of the between study variability, σ_μ, on the accuracy of the estimate of the treatment effect and the posterior SD in Figure 3. The baseline odds for the reference treatment is centered at zero. As σ_μ increases, the accuracy of the estimates obtained by including the single arm evidence decreases to a point where including single arm evidence produces less accurate estimates than the RCT only model. The estimates are more accurate when the baseline odds are close together for each study, as the information is more accurate in the model. However, as the difference increases noise is added into the model by assuming that treatment arms in the matched studies are exchangeable, when in reality they come from different distributions. The plug-in estimator model becomes worse than the pooled model when σ_μ is large. The MAE will, of course, be larger for the treatments that are not in any RCTs. The error for treatments 6 and 7 alone when matching by covariate in the pooled model is between 0.57-1.52 as σ_μ varies from 0 to 1.18.

The crossing point is of particular interest as this is the decision point of whether to include the matched evidence or not. As we see from Figure 3, the crossing point occurs at a higher point for the posterior SD than for the MAE. This implies that at this point we may think that the matched evidence is beneficial, when it is actually harmful. In Table 2 we estimate the between study variability for the HCV infection network, as detailed in Section 4, and a number of publicly

available datasets with binary outcomes from the statistical software package, *R* ([37]). The estimates of σ_μ are of the order of, if not greater than the crossing point in the graph, indicating that matched evidence may often lead to an increase in bias. It may be worth noting that the dataset with the most objective outcome of mortality, i.e. the thrombolytic dataset has one of the lowest between study variability.

Table 2: Between study variability in example networks

Dataset	$\hat{\sigma_\mu}$	**Source**	**Outcome Description**
Thrombolytic	0.36	gemtc in *R*	Mortality after 30-35 days.
Hepatitis C	0.39	details in Section 4	SVR after 12 or 24 weeks.
Certolizumab	0.35	gemtc in *R*	Improvement of at least 50% on the American College of Rheumatology scale (ACR50) at 6 Months.
Smoking	0.59	pcnetmeta in *R*	Successful cessation of smoking at 6-12 months.
Depression	0.61	gemtc in *R*	Reduction of at least 50% from the baseline score on the HAM-D or MADRS at week 8 (or, if not available, another time between week 6 and 12).

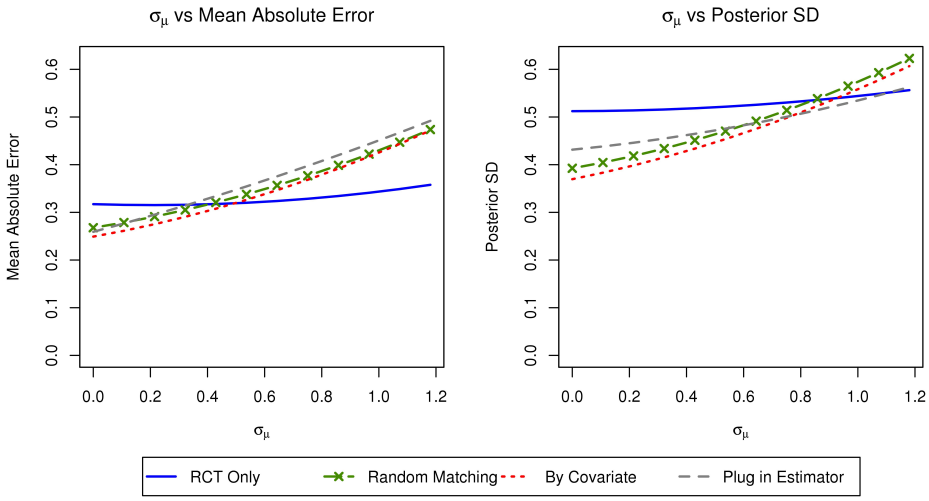


Figure 3: Effect of between study variability, σ_μ, on the MAE and posterior standard deviation. Covariate effect, β=-1.04. The extreme left point on the graph shows the scenario where the study effect is set to zero for every study. The variability between the studies increases with the horizontal axis.

Figure 4 examines the danger of incorrectly assuming that the single arm studies come from the same distribution as the RCTs. We see that the estimates and posterior SDs increase slightly as the bias increases.

3.2.3 *Hierarchical Model*

We now look at the hierarchical model, which includes a between study design effect. We exclude the plug-in estimator model from these results as this model would not be affected by the parameters that are varied in this section. Figure 5 shows how varying the upper bound of the prior on the

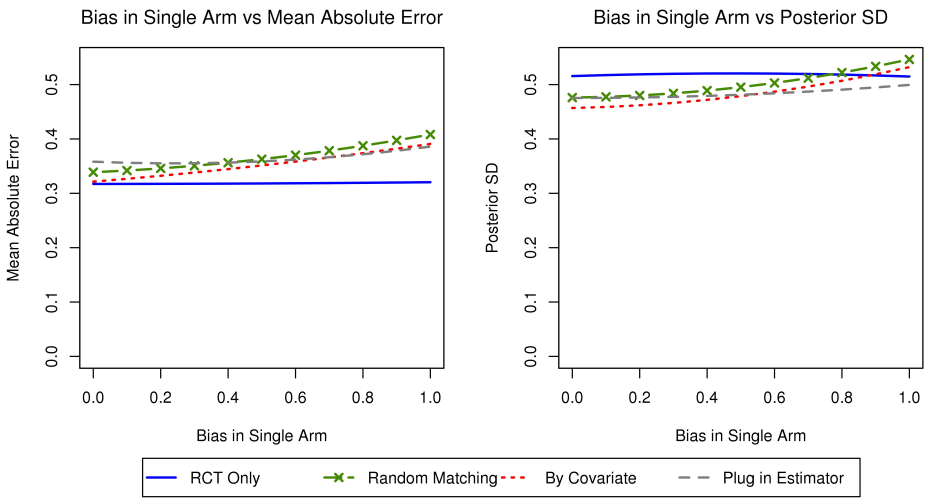


Figure 4: Effect of varying the bias in the single arm trials on the MAE and posterior standard deviation. Between study variability, σ_μ=0.59, covariate effect, β=-1.04. The study effect in the RCTs comes from N(0, 0.59). The value on the x-axis indicates the mean of the single arm trials. Hence, at the extreme left point both the single arms and RCTs come from the same distribution.

between study design effect affects the MAE and the posterior SD. The prior on the between study design effect is given by Unif$(0, z)$ where z varies along the x-axis. The RCT only model has the same estimate at each point as there is only one study type in this model, therefore we have used the average value over all simulations for this line.

As the upper bound of the prior on the between study design variance increases, the MAE and posterior SD also increases. There is a fourth line on the graph corresponding to 12 RCT studies, where we artificially group eight on one side of the hierarchical model, and four on the other side. Here we see the same trend as before, with the posterior SD increasing as the prior on the between study design σ_{des} increases. However, this time it happens to a lesser extent. The increase in posterior SD when including the extra RCTs is solely due to the prior. Any extra increase for the matching methods is due to actual differences between the study types.

Figure 6 shows the effect that down-weighting matched evidence, ω, has on the accuracy of our model's estimate of the treatment effect and the posterior SD. Again, the RCT only line is the average value over all simulations, as there is no weight on the matched evidence. The "RCT By Omega" line shows ω acting on the RCT evidence instead of the (randomly) matched evidence. Decreasing the weight of the high quality RCT evidence generally gives less accurate estimates and larger posterior SDs. However, decreasing the weight of the matched evidence actually gives a concave downwards shape. The MAE and the posterior SD is smallest when $\omega = 0.1$, i.e., the smallest weighting in the simulation study. However, weighting the matched evidence fully appears to be preferable to, or at least as good as, some of the values for ω in the centre of the graph.

4. Example: Hepatitis C Virus - Treatment Naive Patients - NMA

As our research was motivated by a network of treatment regimens for the treatment of HCV infection, we loosely based the simulation study on this network. Chronic HCV infection is a global health burden of major concern. A number of treatment combinations are currently licensed for genotype 1 (GT1) HCV infection. At the time of this systematic review there was a wealth of clinical trial evidence available that compared single regimens in terms of treatment duration and

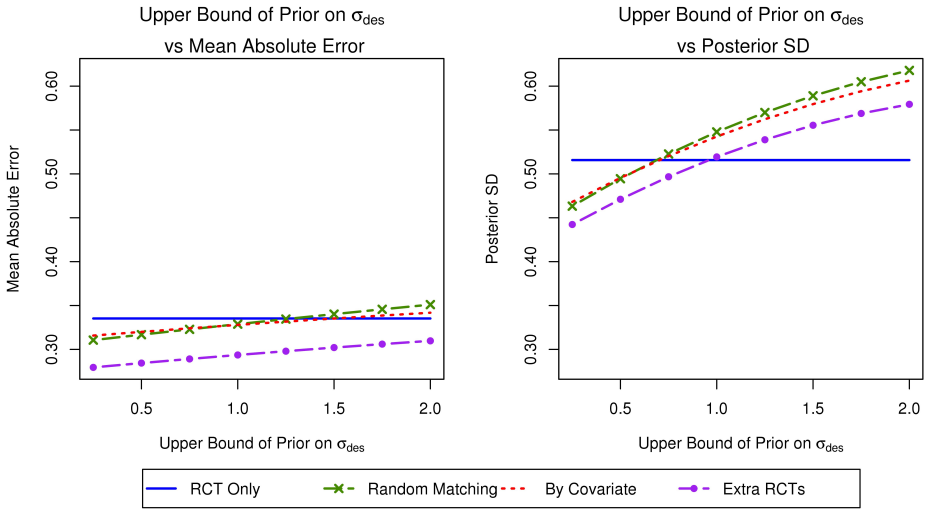


Figure 5: Effect of the prior on the between study design effect (σ_{des}) on the MAE and posterior standard deviation. Between study variability, σ_μ=0.59, covariate effect, β=-1.04. The horizontal axis shows the upper bound for the prior on the between study design effect unif(0, z). Taking two standard deviations for the largest value of an upper bound of 2 corresponds to (0.02, 0.98) on the probability scale.

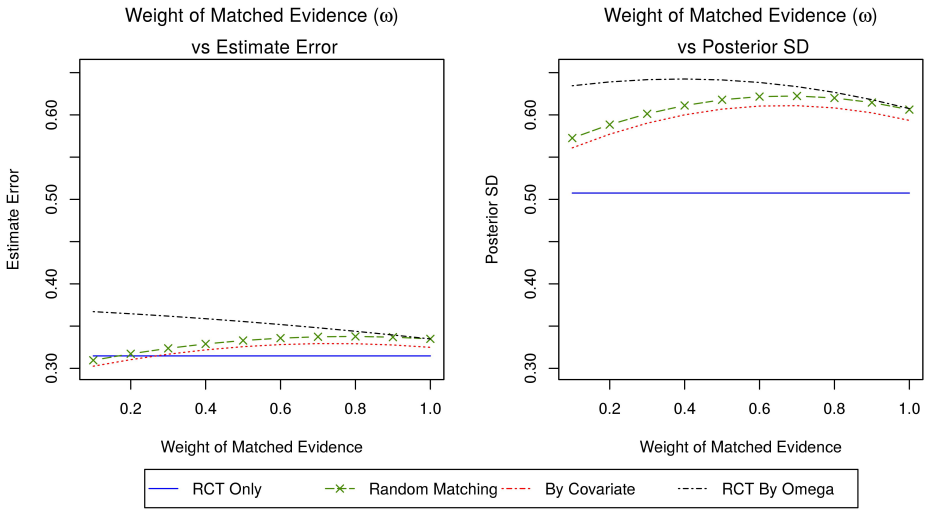


Figure 6: Effect of ω on the MAE and posterior standard deviation. Between study variability, σ_μ=0.59, covariate effect, β=-1.04, prior on between study design effect, $\sigma_{des} \sim$ unif(0, 2).

with or without the addition of ribavirin. However, head-to-head comparative trials between all licensed regimens for GT1 infection were unavailable.

There has been a major change in the way that newer HCV infection treatments have been formulated[22]. There has been a move away from non-specific anti-viral therapies, which had relatively low levels of cure rates, to antiviral combination therapies that directly target replication of the virus, with the ability to significantly enhance cure rates. While RCTs are the most appropriate method to directly assess the relative efficacy of all regimens from a methodological perspective, RCTs comparing newer treatment regimens to older (and most probably inferior) treatments may not be appropriate from an ethical perspective. Therefore, most of the evidence available on the newer HCV infection treatment regimens is disconnected from the older network, and in fact comes in the form of single arm evidence. We therefore apply the techniques discussed in Section 2 to indirectly estimate the relative treatment effect of licensed regimens for the treatment of GT1, by

Table 3: List of treatment regimens with abbreviations

Abbreviation	**Treatment Regimen**
PR	Pegylated-interferon and ribavirin
DCV/PR	Daclatasvir (+ pegylated interferon and ribavirin)
BOC/PR	Boceprevir (+ pegylated interferon and ribavirin)
SIM/PR	Simeprevir (+ pegylated interferon and ribavirin)
TEL/PR	Telaprevir (+ pegylated interferon and ribavirin)
SOF/PR	Sofosbuvir (+ pegylated interferon and ribavirin)
PrOD±RBV	Paritaprevir boosted with ritonavir, ombitasvir and dasabuvir (with or without ribavirin)
SOF/LDV±RBV	Sofosbuvir and ledipasvir (with or without ribavirin)
DCV/SOF±RBV	Daclatasvir + Sofosbuvir (with or without ribavirin)
SIM/SOF±RBV	Simeprevir and sofosbuvir (with or without ribavirin)
SOF/RBV	Sofosbuvir (+ ribavirin)

including single arm trials, in treatment naive patients with chronic GT1 HCV infection.

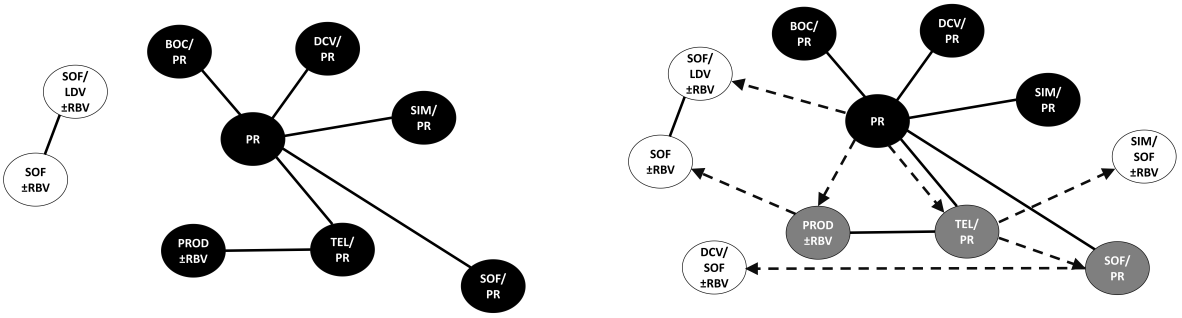
4.1 *Methods*

A systematic review was conducted in accordance with the criteria of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses group (PRISMA)[38] on 6th May 2015, and repeated on 24th November 2015 and on 14th May 2016. These cut off dates are based on the needs of informing the clinical programme in Ireland. We identified 13 RCTs that looked at two interventions[39–51] and 18 single arm trials with no comparator. There were 11 different regimens in our network. In total the single arm studies examined seven different treatment regimens. We chose one arm with no comparator from each treatment regimen to use in this example[52–58]. Where possible we chose single arms that had full information on the covariates of interest. The full list of regimens is described in Table 3. The outcome of interest was a binary outcome, Sustained Virological Response (SVR).

We first ran an analysis which investigated the 13 RCT studies. We then matched the single-arms to an arm of another study to act as the comparator regimen. We then re-ran the meta-analysis using these matched arms. We used the pooled model, four hierarchical models, and the plug-in estimator model. The first hierarchical model gave equal weight to both study types, the other three hierarchical model provided sensitivity analyses by down-weighting the matched estimate by multiplying the matched precision by $\omega = 0.7$, $\omega = 0.4$ and $\omega = 0.1$.

The studies were matched according to the proportion of patients that were cirrhotic, had genotype 1a, and had viral load >800,000 IU/ml at baseline. These covariates were chosen, because according to clinical expert opinion, they were likely to influence SVR, and because these were reported in the majority of trials. Each single arm study was compared to every individual arm (including other single arm studies) and the differences in their baseline characteristics were determined. As we had more than one covariate to use for matching we added the difference in the three covariates together. The arm with the smallest difference in baseline characteristics was then chosen as the matched arm. Not all arms had information on each of the three characteristics. The best match was chosen from the pool of arms that had at least the same amount of information as the single arm trial. A network diagram of the RCT only network and a network including the single-comparator studies is shown in figure 7. Note that for the plug-in estimator model the closest trial was chosen instead of the closest arm, and therefore a network diagram comparing

matched treatments is not applicable for this model.



(a) RCT only

(b) Including matched evidence when matching to arms

Figure 7: Network Diagram. Black nodes denote treatments in RCT trials only, white nodes denote treatments in single arm trials only, and gray nodes denote treatments in both single arm trials and RCT trials. Black solid lines represent RCT connections and dashed lines represent matched connections. The treatment requiring the match has the arrow pointing towards it. Although SOF/LDV±RBV and SOF/RBV are in an RCT they are treated as only being in a single arm trial as the RCT cannot be included as it is not connected to the rest of the network.

4.1.1 *Random Matching Arms - Sensitivity Analysis*

It is possible that the results of the analysis could change based on the choice of matched arms. As there are millions of combinations of arms, it is not possible to test all of them. Therefore, we ran over 1 000 simulations to check matching arms at random in order to assess the sensitivity to the choice of match.

4.1.2 *Removing RCT arms*

In order to assess the accuracy of matching on the Hepatitis C network we investigated how well our method would work on RCTs, which we turned into single arm trials by removing arms. We found an estimate of the efficacy of each drug by using only the RCT evidence. We excluded those that did not have full information for the three covariates and were therefore left with nine RCTs and four treatments: PR, TEL/PR, BOC/PR and SIM/PR. We then selected studies at random and removed all the arms except one. We then matched these new "one-arm" studies to the remaining studies in the network. We compared our results to the results of the reduced dataset with the "one-armed" trials excluded to see which was closer to the estimate with nine trials. We compared these results making a varying amount of studies into single arm trials, from one to six. In all cases we ensured that there was at least one study with each treatment remaining in an RCT so that we were comparing like with like.

4.2 *Results*

Table 4 shows the mean and posterior SD of the log odds ratios (LOR) for each treatment regimen versus PR (standard of care). For treatment regimens which have both RCT and matched evidence (TEL/PR, SOF/PR, and PrOD±RBV) including the matched evidence decreases the posterior SD of the LOR in all cases. The hierarchical model generally results in higher posterior SDs than

Table 4: Log Odds Ratio versus PR

	RCT		Pooled		Hierarchical		Matched Hier Down-weighted $\omega = 0.7$		Matched Hier Down-weighted $\omega = 0.4$		Matched Hier Down-weighted $\omega = 0.1$		Mu From Matched	
Regimen	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**	**Mean**	**Posterior SD**
DCV/PR	2.150	0.898	2.095	0.889	2.181	0.909	2.210	0.934	2.231	0.896	2.211	0.867	2.092	0.891
BOC/PR	1.104	0.232	1.102	0.256	1.107	0.228	1.107	0.218	1.110	0.225	1.111	0.227	1.106	0.224
SIM/PR	1.166	0.222	1.149	0.241	1.157	0.219	1.155	0.211	1.166	0.224	1.162	0.214	1.159	0.217
TEL/PR	1.064	0.208	1.123	0.202	0.928	0.571	0.973	0.549	0.938	0.585	1.025	0.433	1.115	0.176
SOF/PR	1.639	0.648	1.828	0.395	1.496	0.744	1.556	0.734	1.499	0.759	1.625	0.672	2.421	0.326
PrOD±RBV	3.243	0.613	3.560	0.437	3.115	0.794	3.152	0.784	3.087	0.794	3.249	0.693	3.827	0.438
SOF/LDV±RBV			3.944	0.601	3.081	1.161	3.247	1.161	3.172	1.188	3.295	1.116	4.149	0.645
DCV/SOF±RBV			1.898	0.945	1.520	1.320	1.615	1.362	1.587	1.393	1.775	1.417	2.927	1.023
SIM/SOF±RBV			2.757	0.658	2.121	1.215	2.201	1.236	2.163	1.220	2.357	1.231	2.999	0.623
SOF/RBV			0.861	0.685	0.623	1.160	0.708	1.207	0.591	1.172	0.807	1.227	1.120	0.539

the pooled model. This may be explained by the additional prior variance on the study effect. However, for all treatment regimens with RCT evidence only (DCV/PR, BOC/PR and SIM/PR) the adjusted hierarchical model with matched evidence down-weighted has the smallest posterior SD. The pooled model generally gives a more extreme LOR than the hierarchical model, since in the latter case the summary effect is shrunken toward zero. SUrface under the Cumulative RAnking curve (SUCRA) scores[59] are shown in the supplementary material.

4.2.1 *Random Matching Arms - Sensitivity Analysis*

Figure 8 shows the distribution of the LOR for each of treatment regimen versus PR. This can identify how sensitive treatments are to the choice of match. We see that the treatment regimens for which RCTs are available are quite small. However, the treatments regimens for which only matched evidence is available are much larger. We can see that on average newer treatments have the highest LOR (PrOD±RBV, which has both evidence types and SOF/LDV±RBV, DCV/SOF±RBV, and SIM/SOF±RBV which all have single arm evidence only). This highlights the importance of including these treatments in an NMA as they are likely to be better than the older treatment regimens, which are included in RCTs. It is important to note here that most estimates from the chosen match give a smaller LOR than average. In particular we compare the best match hierarchical model (denoted by a red triangle) with the distribution shown, as these are the same models. This reassuringly highlights that we are at least being conservative with our estimates of the single arm treatments.

4.2.2 *Removing RCT arms*

As we can see in Figure 9, in the hepatitis C network if we change three or fewer RCTs to single arm trials the reduced model is slightly closer to the full network than using our matched methods, on average. However, if we change five or six RCTs into single arm trials then all methods gives results that are much closer to the full network than the results we obtain from the reduced dataset. We also see that the estimate of the reduced dataset is less precise than the pooled model, and thus the reduced model has the same or higher posterior SD than the plug-in estimator model. A "U-shaped" curved is noticeable for the hierarchical model for the posterior SD. The reason for the high uncertainty when we only match one or two studies is because we have very little evidence on the matched side of the hierarchical model. From this analysis the pooled model would be the most preferable.

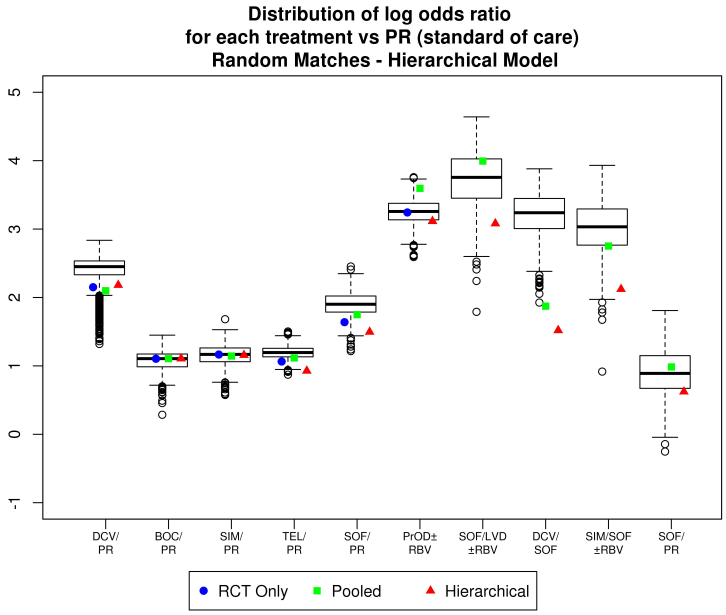


Figure 8: Distribution of the log odds ratio using randomly matching arms compared with results from RCT only and the best match chosen by the equal weights method.

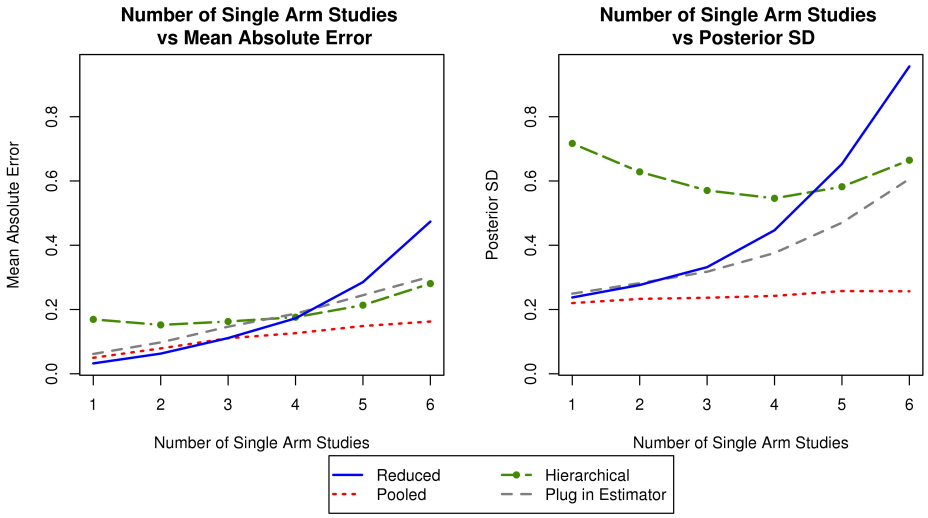


Figure 9: We use only one arm from a number of RCTs (as indicated on the x-axis). Results from the reduced RCT network or network obtained by matching the new "single arm" studies are compared to our best estimate of the treatment effect, i.e. the full nine study network.

5. Discussion

5.1 *Summary*

If there was sufficient RCT evidence there would be no need to include single arm evidence as it is of lesser quality. However, sometimes only single arm evidence is available. Single arm evidence is currently being used in HTA, so it is important that we find the most suitable method for including such single arm evidence in an NMA, and that we understand the benefits and limitations of this type of evidence. Figure 10 shows a schematic representation of the association between some of the parameters considered in this paper and the MAE and posterior SD.

Other factors such as the covariate effect, the treatment effect and the size of the RCTs can

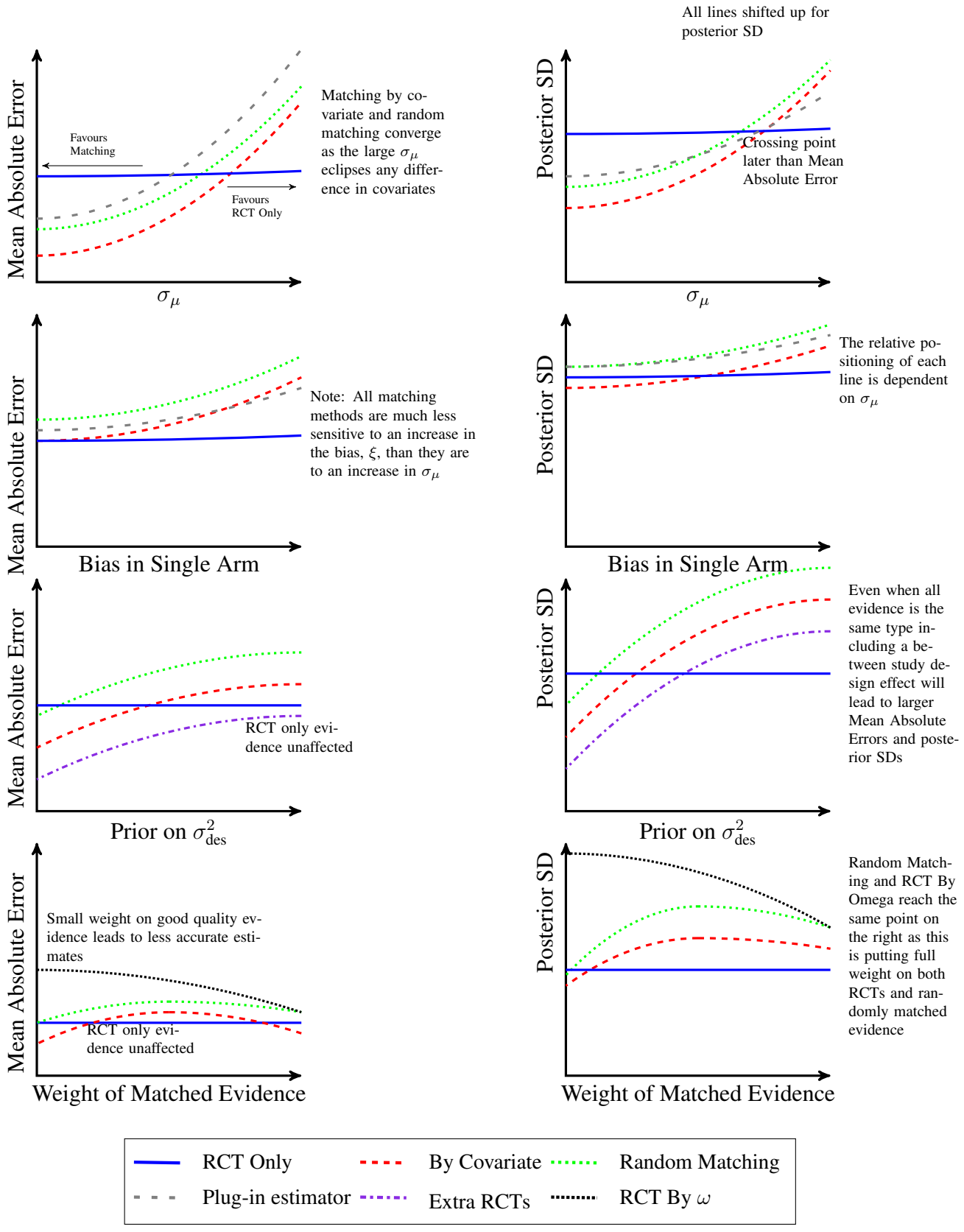


Figure 10: Schematic representation of the association between the parameters and the mean absolute error and posterior SD based on the results from the simulation study. Note that location and size of effects may depend on other parameters as described in the text.

also affect the appropriateness of including matched evidence. For these we tried to choose the parameters in the simulation study so that they were as realistic as possible, and if anything they favoured RCT evidence. In the exploratory analysis we found that matching was most valuable when trials were small. Similarly, in the Hepatitis C example the matching method only showed superiority over the reduced method once four or five studies had been removed.

There is a concern that allowing the inclusion of single arm evidence would disincentivise further RCT research from being undertaken. There is also the question of whether we are willing to accept a particularly biased estimate over no estimate at all. Therefore it makes sense to have the option to down-weight the single arm evidence, so that it does not have the same impact as RCT evidence.

5.2 *Recommendations*

We recommend considering single-arm evidence when the variability between studies is small and when we can find an appropriate match. We can attempt to quantify this variability by examining the baseline effect across the RCTs. As the discrepancies between the studies increase, the incorrect assumption of exchangeablity between matched arms leads to a decreased accuracy and precision in the estimates. This form of evidence has a high possibility of being biased and should be used with caution.

When undertaking single-arm matching there are a number of sensitivity analyses that should be carried out:

1. Quantify the variability between RCTs;
2. Compare the results using the best match to the results using random matches;
3. Adjust the weight of the matched evidence in the hierarchical model by increasing the variance inflation.

It is important to emphasise that any method for incorporating single arm evidence cannot replace the unbiased approach of an RCT. Given the large potential financial gain for a pharmaceutical company arising from a positive HTA recommendation, there is a clear incentive to ensure that treatments look as effective as possible. Single arm trials are much more exposed to bias or manipulation than RCTs[60]. Therefore, it is imperative that we bear this in mind when making decisions based on single arm evidence. While aggregate level matching may be helpful in estimating treatment rankings when no other evidence is available, we would not consider this type of evidence to be convincing enough to be the primary source of evidence when making reimbursement decisions. Although there are advantages in using single-arm evidence, the burden of proof must be on the pharmaceutical company to demonstrate the benefits of their treatments using methods that are as free from bias as possible.

5.3 *Limitations and Extensions*

In our simulations we made many assumptions about the size, shape and structure of the network. The prior in the hierarchical model may be too vague for sparse networks, especially considering that it may be quite difficult to estimate σ_{des} given that we have only two study types. Therefore, it may be worth considering an informative prior distribution formed by expert opinion as suggested by Efthimiou et al[20]. Although our worked example only uses binary covariates, this method can easily be applied to continuous covariates by using the mean in each trial.

In this paper we choose the closest match for each treatment regimen. However, there may be other studies or arms that are nearly as similar as the chosen match. On the other hand, there may be some treatments regimens for which all potential matches are very dissimilar to the single arm

trial. In this case we may want to consider excluding this treatment regimen from the analysis. Criteria for assessing a sufficiently similar match is discussed in Schmitz et al[29].

It would also be worth checking if the best match is usually more conservative than a random match in various networks. If so, it may indicate that single arm trials are more likely to give better results (perhaps because of healthier patient recruitment) than RCTs and reinforce the point that single arm evidence should be used with caution and covariates should be matched as much as possible.

5.4 *Conclusion*

There is increasing motivation to use single arm evidence where available, as pharmaceutical companies increasingly use this type of evidence to obtain regulatory approval. Methods such as those explored in this paper are already being implemented in clinical practice, as seen in Jaff et al[28]. There is a high risk of bias and considerable uncertainty arising from incorporating single arm evidence into an NMA. Therefore it is imperative that the methods used are transparent, backed up by systematic investigation, and a clear indication for how much bias could potentially be introduced is provided.

6. Acknowledgments

This research was funded by the Health Research Board, Ireland, Grant Number HRB RL2013/4 (PI: Cathal Walsh). Some of the computations in this work made use of the statistics computing cluster which was supported by the STATICA project which was funded by the Principal Investigator programme of Science Foundation Ireland, Grant number 08/IN.1/I1879 (PI: Simon P. Wilson).

References

1. Lumley Thomas. Network meta-analysis for indirect treatment comparisons *Statistics in medicine.* 2002;21:2313–2324.
2. Lu Guobing, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons *Statistics in medicine.* 2004;23:3105–3124.
3. Cooper Nicola J, Peters Jaime, Lai Monica CW, et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value in Health.* 2011;14:371–380.
4. Senn Stephen, Gavini Francois, Magrez David, Scheen André. Issues in performing a network meta-analysis *Statistical Methods in Medical Research.* 2013;22:169–189.
5. Salanti Georgia, Higgins Julian PT, Ades AE, Ioannidis John PA. Evaluation of networks of randomized trials *Statistical methods in medical research.* 2008;17:279–301.
6. Song Fujian, Clark Allan, Bachmann Max O, Maas Jim. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons *BMC Medical Research Methodology.* 2012;12:1-14.
7. Jansen Jeroen P, Naci Huseyin. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers *BMC medicine.* 2013;11:1.
8. Welton Nicky J, Sutton Alexander J, , Cooper Nicola, Abrams Keith R, Ades AE. *Evidence synthesis for decision making in healthcare*. John Wiley & Sons 2012.
9. Sterne Jonathan AC, Hernán Miguel A, Reeves Barnaby C, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions *British Medical Journal.* 2016;355:i4919.

10. Cameron Chris, Fireman Bruce, Hutton Brian, et al. Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments: challenges and opportunities *Systematic reviews.* 2015;4:147.
11. Valentine Jeffrey C, Thompson Simon G. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions *Research synthesis methods.* 2013;4:26–35.
12. Griffiths Elizabeth A, Macaulay Richard, Vadlamudi Nirma K, Uddin Jasim, Samuels Ebony R. The Role of Noncomparative Evidence in Health Technology Assessment Decisions *Value in Health.* 2017.
13. Bell Helen, Wailoo Allan J, Hernandez Monica, et al. The use of real world data for the estimation of treatment effects in NICE decision making. Report by the Decision Support Unit, ScHARR, University of Sheffield 2016.
14. Signorovitch James E, Wu Eric Q, Andrew P Yu, et al. Comparative effectiveness without head-to-head trials *Pharmacoeconomics.* 2010;28:935–945.
15. Signorovitch James E, Sikirica Vanja, Erder M Haim, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research *Value in Health.* 2012;15:940–947.
16. Caro J Jaime, Ishak K Jack. No head-to-head trial? Simulate the missing arms *Pharmacoeconomics.* 2010;28:957–967.
17. Ishak KJ, Proskorovsky I, Benedict A, Chen C. Simulated Treatment Comparisons–An Alternative Approach to Indirect Comparison When Standard Methods Are Not Feasible or Appropriate *Value in Health.* 2013;16:A615.
18. Sutton Alex J, Abrams Keith R. Bayesian methods in meta-analysis and evidence synthesis *Statistical methods in medical research.* 2001;10:277–303.
19. Schmitz Susanne, Adams Roisin, Walsh Cathal. Incorporating data from various trial designs into a mixed treatment comparison model *Statistics in medicine.* 2013;32:2935–2949.
20. Efthimiou Orestis, Mavridis Dimitris, Debray Thomas, et al. Combining randomized and non-randomized evidence in network meta-analysis *Statistics in medicine.* 2017;36:1210–1226.
21. Thom Howard HZ, Capkun Gorana, Cerulli Annamaria, Nixon Richard M, Howard Luke S. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension *BMC medical research methodology.* 2015;12:15-34.
22. Goring SM, Gustafson P, Liu Y, Saab S, Cline SK, Platt RW. Disconnected by design: analytic approach in treatment networks having no common comparator *Research Synthesis Methods.* 2016.
23. Hong Hwanhee, Chu Haitao, Zhang Jing, Carlin Bradley P. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons *Research synthesis methods.* 2016;7:6-22.
24. Hong H, Chu H, Zhang J, Carlin Bradley P. Rejoinder to the discussion of "A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons" by S. Dias and AE Ades *Research synthesis methods.* 2016;7:29-33.
25. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data *Research synthesis methods.* 2016;7:23-28.
26. Phillippo David, Ades Tony, Dias Sofia, Palmer Stephen, Abrams Keith R, Welton Nicky. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE 2016.

27. Phillippo David M, Ades Anthony E, Dias Sofia, Palmer Stephen, Abrams Keith R, Welton Nicky J. Methods for population-adjusted indirect comparisons in health technology appraisal *Medical Decision Making.* 2018;38:200–211.
28. Jaff Michael R, Nelson Teresa, Ferko Nicole, Martinson Melissa, Anderson Louise H, Hollmann Sarah. Endovascular Interventions for Femoropopliteal Peripheral Artery Disease: A Network Meta-Analysis of Current Technologies *Journal of Vascular and Interventional Radiology.* 2017.
29. Schmitz Susanne, Maguire Áine, Morris James, et al. The use of single armed observational data to closing the gap in otherwise disconnected evidence networks: a network meta-analysis in multiple myeloma *BMC medical research methodology.* 2018;18:66.
30. Dias Sofia, Sutton Alex J, Ades AE, Welton Nicky J. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials *Medical Decision Making.* 2013;33:607–617.
31. Kass Robert E, Wasserman Larry. The selection of prior distributions by formal rules *Journal of the American Statistical Association.* 1996;91:1343–1370.
32. Leahy Joy, O'Leary Aisling, Afdhal Nezam, et al. The Impact of Individual Patient Data in a Network Meta Analysis: An investigation into parameter estimation and model selection *Research Synthesis Methods.* 2018;9:441-469.
33. Lambert Paul C, Sutton Alex J, Burton Paul R, Abrams Keith R, Jones David R. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS *Statistics in medicine.* 2005;24:2401–2428.
34. Borenstein Michael, Hedges Larry V, Higgins Julian, Rothstein Hannah R. *Meta-Regression*. Wiley Online Library 2009.
35. Spiegelhalter David, Thomas Andrew, Best Nicky, Lunn Dave. OpenBUGS User Manual 2014.
36. Gelman Andrew, Rubin Donald B. Inference from iterative simulation using multiple sequences *Statistical science.* 1992;7:457–472.
37. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical ComputingVienna, Austria 2013.
38. Moher David, Liberati Alessandro, Tetzlaff Jennifer, Altman Douglas G. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement *Annals of internal medicine.* 2009;151:264–269.
39. Jacobson Ira M, McHutchison John G, Dusheiko Geoffrey, et al. Telaprevir for previously untreated chronic hepatitis C virus infection *New England Journal of Medicine.* 2011;364:2405–2416.
40. Hézode Christophe, Forestier Nicole, Dusheiko Geoffrey, et al. Telaprevir and peginterferon with or without ribavirin for chronic HCV infection *New England Journal of Medicine.* 2009;360:1839–1850.
41. McHutchison John G, Everson Gregory T, Gordon Stuart C, et al. Telaprevir with peginterferon and ribavirin for chronic HCV genotype 1 infection *New England Journal of Medicine.* 2009;360:1827–1838.
42. Poordad Fred, McCone Jr Jonathan, Bacon Bruce R, et al. Boceprevir for untreated chronic HCV genotype 1 infection *New England Journal of Medicine.* 2011;364:1195–1206.
43. Kwo Paul Y, Lawitz Eric J, McCone Jonathan, et al. Efficacy of boceprevir, an NS3 protease inhibitor, in combination with peginterferon alfa-2b and ribavirin in treatment-naive patients with genotype 1 hepatitis C infection (SPRINT-1): an open-label, randomised, multicentre phase 2 trial *The Lancet.* 2010;376:705–716.

44. Kumada Hiromitsu, Toyota Joji, Okanoue Takeshi, Chayama Kazuaki, Tsubouchi Hirohito, Hayashi Norio. Telaprevir with peginterferon and ribavirin for treatment-naive patients chronically infected with HCV of genotype 1 in Japan *Journal of hepatology.* 2012;56:78–84.
45. Jacobson Ira M, Dore Gregory J, Foster Graham R, et al. Simeprevir with pegylated interferon alfa 2a plus ribavirin in treatment-naive patients with chronic hepatitis C virus genotype 1 infection (QUEST-1): a phase 3, randomised, double-blind, placebo-controlled trial *The Lancet.* 2014;384:403–413.
46. Manns Michael, Marcellin Patrick, Poordad Fred, et al. Simeprevir with pegylated interferon alfa 2a or 2b plus ribavirin in treatment-naive patients with chronic hepatitis C virus genotype 1 infection (QUEST-2): a randomised, double-blind, placebo-controlled phase 3 trial *The Lancet.* 2014;384:414–426.
47. Fried Michael W, Buti Maria, Dore Gregory J, et al. Once-daily simeprevir (TMC435) with pegylated interferon and ribavirin in treatment-naïve genotype 1 hepatitis C: The randomized PILLAR study *Hepatology.* 2013;58:1918–1929.
48. Pol Stanislas, Ghalib Reem H, Rustgi Vinod K, et al. Daclatasvir for previously untreated chronic hepatitis C genotype-1 infection: a randomised, parallel-group, double-blind, placebo-controlled, dose-finding, phase 2a trial *The Lancet infectious diseases.* 2012;12:671–677.
49. Lawitz Eric, Lalezari Jay P, Hassanein Tarek, et al. Sofosbuvir in combination with peginterferon alfa-2a and ribavirin for non-cirrhotic, treatment-naive patients with genotypes 1, 2, and 3 hepatitis C infection: a randomised, double-blind, phase 2 trial *The Lancet infectious diseases.* 2013;13:401–408.
50. Gane Edward J, Stedman Catherine A, Hyland Robert H, et al. Efficacy of nucleotide polymerase inhibitor sofosbuvir plus the NS5A inhibitor ledipasvir or the NS5B non-nucleoside inhibitor GS-9669 against HCV genotype 1 infection *Gastroenterology.* 2014;146:736–743.
51. Dore Gregory J, Conway Brian, Luo Yan, et al. Efficacy and safety of ombitasvir/paritaprevir/r and dasabuvir compared to IFN-containing regimens in genotype 1 HCV patients: The MALACHITE-I/II trials *Journal of hepatology.* 2016;64:19–28.
52. Sherman Kenneth E, Flamm Steven L, Afdhal Nezam H, et al. Response-guided telaprevir combination treatment for hepatitis C virus infection *New England Journal of Medicine.* 2011;365:1014–1024.
53. Feld Jordan J, Kowdley Kris V, Coakley Eoin, et al. Treatment of HCV with ABT-450/r–ombitasvir and dasabuvir with ribavirin *New England Journal of Medicine.* 2014;370:1594–1603.
54. Osinusi Anuoluwapo, Meissner Eric G, Lee Yu-Jin, et al. Sofosbuvir and ribavirin for hepatitis C genotype 1 in patients with unfavorable treatment characteristics: a randomized clinical trial *Jama.* 2013;310:804–811.
55. Afdhal Nezam, Zeuzem Stefan, Kwo Paul, et al. Ledipasvir and sofosbuvir for untreated HCV genotype 1 infection *New England Journal of Medicine.* 2014;370:1889–1898.
56. Kwo Paul, Gitlin Norman, Nahass Ronald, et al. A phase-3, randomised, open-label study to evaluate the efficacy and safety of 8 and 12 weeks of Simeprevir (SMV) plus Sofosbuvir (SOF) in treatment-naive and-experienced patients with chronic HCV genotype 1 infection without cirrhosis: Optimist-1 *J Hepatol.* 2015;62:S270.
57. Sulkowski Mark S, Gardiner David F, Rodriguez-Torres Maribel, et al. Daclatasvir plus sofosbuvir for previously treated or untreated chronic HCV infection *New England Journal of Medicine.* 2014;370:211–221.
58. Lawitz Eric, Mangia Alessandra, Wyles David, et al. Sofosbuvir for previously untreated chronic hepatitis C infection *New England Journal of Medicine.* 2013;368:1878–1887.

59. Salanti Georgia, Ades A.E., Ioannidis John P.A.. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial *Journal of Clinical Epidemiology.* 2011;64:163 - 171.
60. Grieve Richard, Abrams Keith, Claxton Karl, et al. Cancer Drugs Fund requires further reform *Bmj.* 2016;354.